Statistical evaluation of diet-microbe associations

Zhang, Xiang; Nieuwdorp, Max; Groen, Albert K; Zwinderman, Aeiko H

Published in:
BMC Microbiology

DOI:
10.1186/s12866-019-1464-0

Publication date:
2019

Zhang et al. BMC Microbiology (2019) 19:90
https://doi.org/10.1186/s12866-019-1464-0

BMC Microbiology

RESEARCH ARTICLE Open Access

Statistical evaluation of diet-microbe associations

Xiang Zhang[1*], Max Nieuwdorp[2], Albert K. Groen[1] and Aeiko H. Zwinderman[3]

Abstract

Background: Statistical evaluation of the association between microbial abundance and dietary variables can be done in various ways. Currently, there is no consensus on which methods are to be preferred in which circumstances. Application of particular methods seems to be based on the tradition of a particular research group, availability of experience with particular software, or depending on the outcomes of the analysis.

Results: We applied four popular methods including edgeR, limma, metagenomeSeq and shotgunFunctionalizeR, to evaluate the association between dietary variables and abundance of microbes. We found large difference in results between the methods. Our simulation studies revealed that no single method was optimal.

Conclusions: We advise researchers to run multiple analyses and focus on the significant findings identified by multiple methods in order to achieve a better control of false discovery rate, although the false discovery rate can still be substantial.

Keywords: Microbiome, Diet, Association, Simulation, Sequencing

Background

With the help of high-throughput sequencing technologies, human microbiota have been profiled and studied extensively [1]. Since diet shapes the composition of human microbiota and influences human health, linking abundance of microbes to dietary variables is a common practice in human microbiome studies [2, 3]. These association studies not only can improve our understanding of the relationships between the human microbiome and nutrient intake, but also may help development of new therapeutic interventions.

Microbiome data are often generated by targeted sequencing of the 16S ribosomal RNA (rRNA) gene, and represented as a frequency matrix giving the number of times each microbe is observed in each sample. In general microbiome data have following features: 1) library sizes can vary by orders of magnitude across samples. 2) microbiome data often have excess zero counts. These zero counts can be due to either biological absence of a microbe, or insufficient sequencing. 3) microbiome data are compositional data, meaning that the obtained counts do not reflect the absolute number of microbes that are present. 4) microbiome data are over-dispersed, characterized as some taxa (e.g., *Bacteroides* and *Lactobacillus* species) are common among samples, many other taxa are present at much lower abundances.

Various statistical methods have been developed for microbiome data analysis, but there are no standard procedures to perform association analyses [4]. Previous benchmark works [5, 6] focused on case-control studies, and revealed that the choice of statistical methods considerably affected outcomes of differential relative abundance tests. Unlike case-control studies, association studies work also on continuous variables. To our best knowledge, the influence of choosing different methods on outcomes of association studies has not been evaluated. To assess the influence, we analyzed the associations between dietary variables and gut microbiota in 1090 individuals from the HELIUS-cohort study (Amsterdam, the Netherlands) [7, 8]. Since the focus of the current work is on robustness of the statistical results rather than biological or epidemiological associations, biological interpretation of diet-microbe associations is out of the scope of this work. We used four methods including those based on Poisson (shotgunFunctionalizeR), negative binomial

* Correspondence: xiang.zhang@amsterdamumc.nl
[1]Department of Experimental Vascular Medicine, Amsterdam University Medical Center, Meibergdreef 9, Amsterdam, The Netherlands
Full list of author information is available at the end of the article

(edgeR), zero-inflated Gaussian (metagenomeSeq) distributions, as well as a weighted linear regression model (voom + limma). We compared the results derived from the four methods and observed large differences. To find out which method we should choose in which circumstances, we ran simulation studies and found that no single method was optimal for all microbiome data sets. We advise researchers to run multiple statistical analyses and focus on the significant findings identified by multiple methods in order to achieve better control of false discovery rate.

Results

Large difference in results between statistical analyses

To evaluate effect of choosing different methods on outcomes in association studies, we performed association analyses between 67 dietary variables and 2073 OTUs derived from 1090 HELIUS participants with four methods. Out of 138,891 association tests, we identified 3535, 20,081, 62,581 and 71,371 associations with FDR below 0.05 in edgeR, voom + limma, metagenomeSeq and shotgunFunctionalizeR, respectively. There were 1296 associations identified to be significant by all the four methods. In addition, there were 14, 3703, 23,666, and 29,327 associations that were identified as significant only by edgeR, voom + limma, metagenomeSeq or shotgunFunctionalizeR (Fig. 1).

16S rRNA microbiome data simulation

After realizing such considerably different results between the methods, we attempted to find out which method we should choose. To this end, we simulated 16S rRNA microbiome data with spiked-in associations between dietary variables and OTUs. We used a published FFQs (food frequency questionnaires) data as a template. To make sure our simulation framework can generate similar microbiome data as real ones, we compared our simulated data to the real HMP (Human Metabolome Project) stool 16S data. Our simulated microbiome data had similar distribution of library sizes and percentage of zeros per OTU, as well as similar mean-variance relationship (Fig. 2). Our template FFQs data contained 214 dietary variables. In each simulation, we used one dietary variable. Therefore, in total we generated 214 simulated 16S rRNA data sets. Each data set contained 1000 subjects and had mean library size

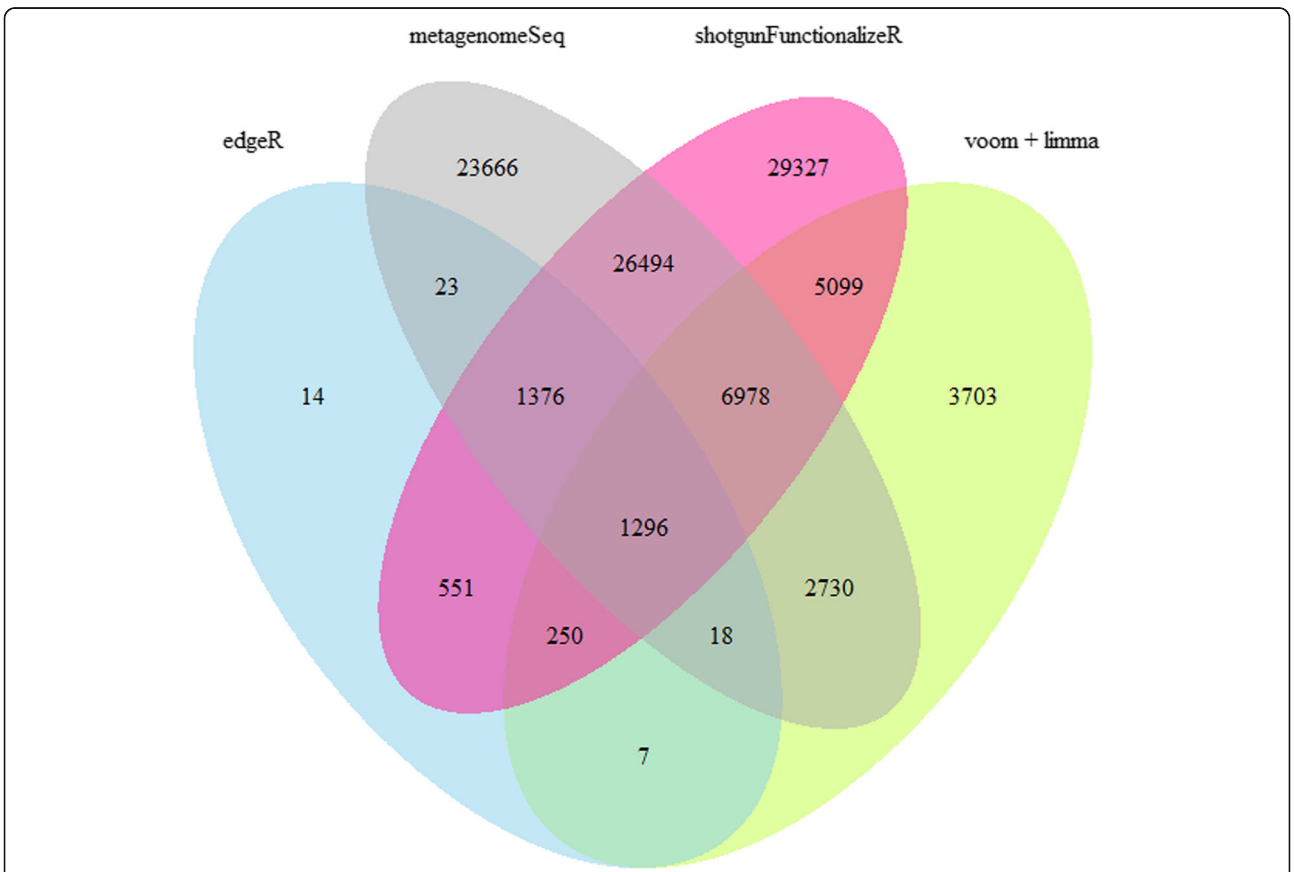

Fig. 1 Venn diagram of significant associations identified by edgeR, voom + limma, metagenomeSeq and shotgunFunctionalizeR based on HELIUS 16S rRNA microbiome and FFQ data

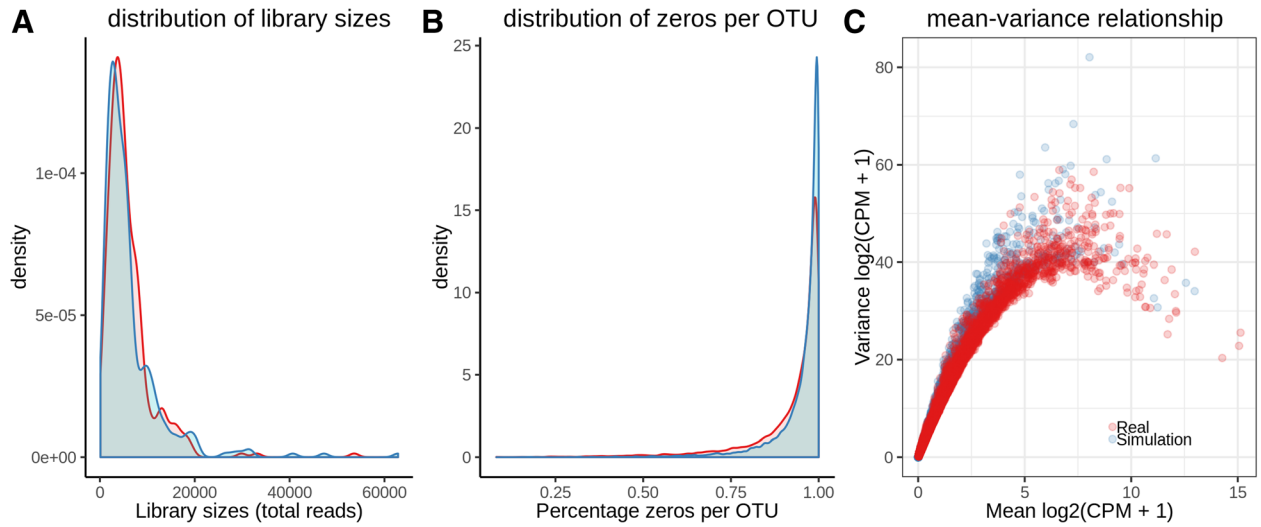

Fig. 2 Comparison of simulated and real Human Microbiome Project stool 16S rRNA data. A: library size distributions B. distribution of percentage of zeros per OTU. C. mean-variance relationship. Every dot represents an OTU

50,000, and the same simulated data set was analyzed by edgeR, voom + limma, metagenomeSeq and shotgunFunctionalizeR. In our simulations, we observed large difference in results between the methods (Fig. 3).

Method comparisons based on simulated data

Overall shotgunFunctionalizeR had both the highest true positive rate and the highest false positive rate (Fig. 4). The median true positive rate of shotgunFunctionalizeR was (0.900), followed by metagenomeSeq (0.800), edgeR (0.624) and limma (0.519). Meanwhile the median false positive rate of shotgunFunctionalizeR, metagenomeSeq, limma, and edgeR were 0.716, 0.388, 0.125 and 0.0898, respectively. Based on the 214 simulations, we identified that the median error probability, defined as the probability that a significant association is false, of shotgunFunctionalizeR, metagenomeSeq, limma and edgeR were 0.439, 0.330, 0.196 and 0.123, respectively (Fig. 5a). Among the 214 simulations, we observed that edgeR had the lowest error probability compared to other methods in 147 simulations, followed by limma (56 simulations) and metagenomeSeq (11 simulations) (Fig. 5b). Furthermore, the error probabilities in different methods were also influenced by the skewness of the distribution of the dietary variables (Fig. 6). In the next step, we identified that 30% simulations in edgeR, 16% simulations in limma, 0.9% simulations in metagenomeSeq and 0% simulation in shotgunFunctionalizeR had error probabilities below 0.05 over the 214 simulations (Fig. 7). However, when we focused on the significant associations that were identified by all four methods (we call them "overlap") in each simulation, we observed that 44% simulations had error probabilities below 0.05 over the 214 simulations (Fig. 7).

Discussion

We learned from these relatively simple analyses that a key issue in the analysis of 16S rRNA microbiome data is the choice of the statistical method. Depending on the choice of statistical method, significant associations between dietary variables and microbial abundances varied dramatically. We observed that shotgunFunctionalizeR produced the largest number of unique significant associations, whereas most of the significant associations identified by edgeR were also identified by other methods. What really puzzled us is the relatively small number of significant associations identified by all methods. We think that such dramatic difference is related to the distribution assumptions as well as the normalization processes implemented in the statistical methods [9]. In this study, shotgunFunctionalizeR and edgeR modeled the unnormalized counts with either Poisson or negative binomial distribution, and coped with uneven library size across samples by including the log(total counts) as the offset. In contrast to shotgunFunctionalizeR and edgeR, limma and metagenomeSeq were based on the Gaussian and zero-inflated Gaussian distributions, requiring transforming discrete counts into continuous quantities. To this end, limma transformed the raw counts into log-cpm (log counts per million), in which unequal library sizes were normalized. In metagenomeSeq, uneven library size was normalized by cumulative sum scaling, and the normalized counts were log2 transformed in order to be incorporated into the zero-inflated Gaussian model.

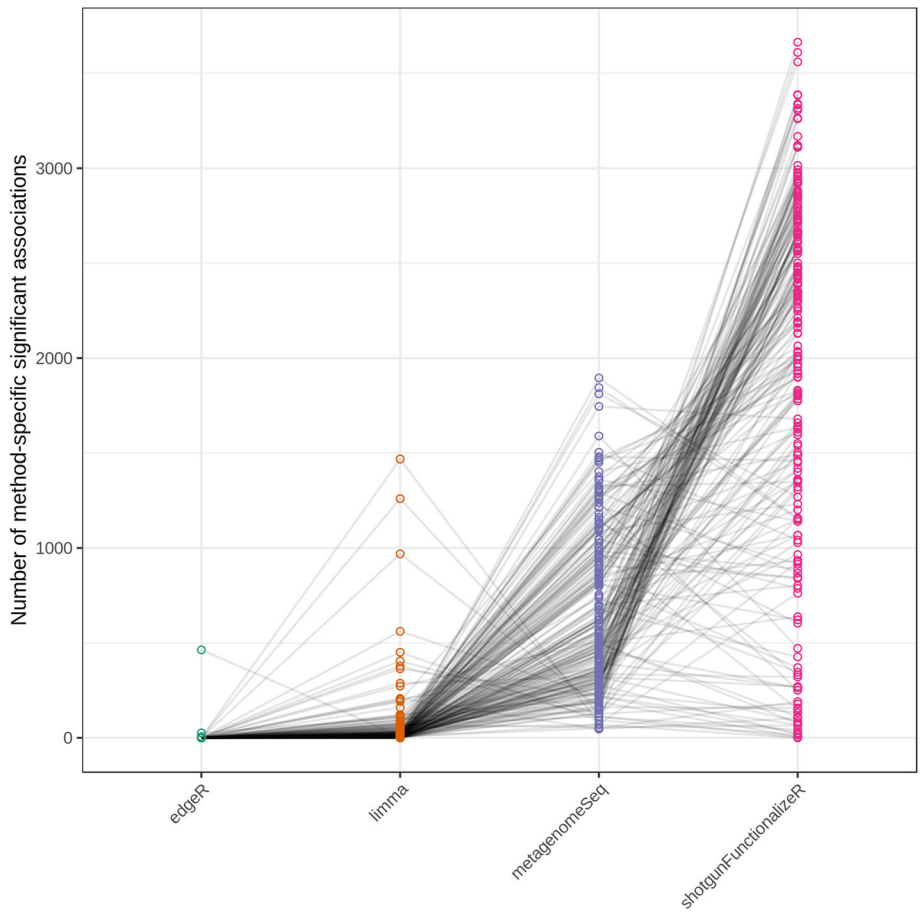

Fig. 3 Every dot represents the number of significant associations identified only by the corresponding method. Each line represents a simulation, in which the same simulated data were analyzed by edgeR, voom + limma, metagenomeSeq and shotgunFunctionalizeR

To find out which method we should choose for association studies, we developed a hierarchical model to simulate 16S rRNA data based on dietary variables with spiked-in associations. By comparing to the real HMP 16S microbiome data, we have shown that our simulation model can simulate realistic 16S rRNA microbiome data. Although in this work we focused on diet-microbe association analyses, our simulation framework can easily be adapted to simulate other scenarios.

Based on our simulation model, we generated a large number of 16S microbiome data sets with sample size 1000 subjects and mean of sequencing depth 50,000. These settings were used to mimic the HELIUS data set. When we analyzed the simulated data sets with edgeR, limma, metagenomeSeq and shotgunFunctionalizeR, we observed again large difference in number of significant associations between the methods. In general, we want our statistical method to detect as many as possible true positives, and as few as possible false positives. From our simulation studies, we learned that overall the most sensitive method (shotgunFunctionalizeR in this case) was likely to be the one with the most false positives. This phenomenon was observed in the differential abundance test scenario as well [5]. Even though we set FDR as 0.05 in all our diet-microbe association analyses, our simulation results showed that control of FDR completely failed in shotgunFunctionalizeR, and rarely achieved in metagenomeSeq. On the other hand, edgeR and limma achieved FDR 0.05 in some cases. In the previous case-control simulations [6], metagenomeSeq and shotgunFunctionalizeR were shown to fail controlling false discovery rate at 0.05. However, edgeR was reported to be able to control false discovery rate at 0.05 [6]. We think this is due to the fact that performing association analyses is more challenging than case-control comparisons because we cannot control both dependent and independent variables. Our further analysis showed that the skewness of the independent variable (e.g. dietary variable) influences the error probabilities in all methods. When the skewness

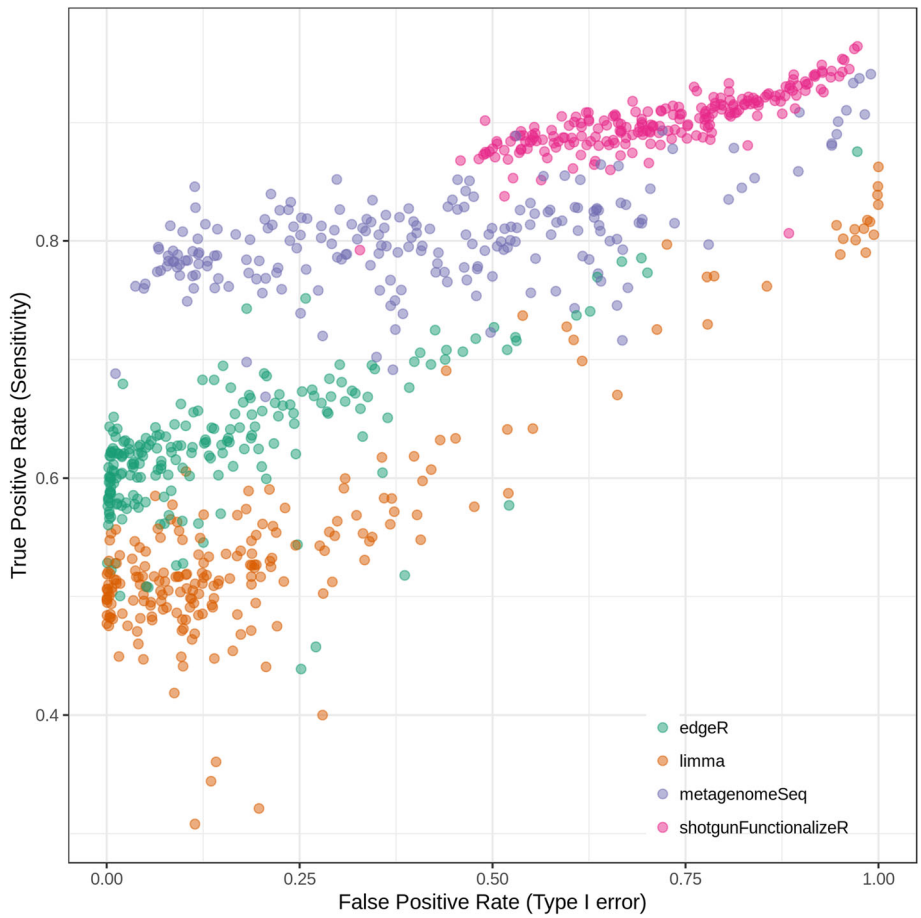

Fig. 4 With each simulated data set, we calculated the performance of every method, in terms of true positive rate and false positive rate. Every dot represents a simulation

of the dietary variable increased, the probability that a significant association is false also increased. When we focused on the significant diet-microbe associations that were identified by all four methods, we observed that more simulations had error probability below 0.05.

Conclusions

In summary, the choice of the statistical method is a key issue in the analysis of 16S rRNA microbiome data. No single method was optimal for diet-microbe association analyses. We recommend researchers to run multiple statistical models and focus on the significant associations identified by multiple methods. In this way, we can improve the controlling of false discovery rate, although the false discovery rate can still be substantial.

Methods

Subjects and HELIUS cohort

Subjects were participants in the HEalthy Life in an Urban Setting (HELIUS) cohort study. This study used a stratified-random sampling approach to include between 2011 and 2015 25,000 inhabitants (18–70 years) from the city of Amsterdam, the Netherlands [7]. Stratification was done on six subgroups with different ethnic origins (African Surinamese, South Asian Surinamese, Ghanaian, Turkish, Moroccan, and Dutch). Subgroups were about equally large.

Dietary intakes assessment

As described previously [10, 11], a subsample of voluntary participants of Dutch, Moroccan, Turkish, South-Asian Surinamese and African Surinamese origin were asked to participate in the HELIUS-Dietary Patterns study, with objective to collect detailed information on their diet. Usual dietary intakes were assessed through the completion of ethnic-specific semi-quantitative food frequency questionnaires (FFQs) with a reference period of 4 weeks. These FFQs were adapted from an existing Dutch FFQ and comprised about 200 items. Food items were collapsed into 73 food groups based on similarity in nutrient profile and culinary use. In

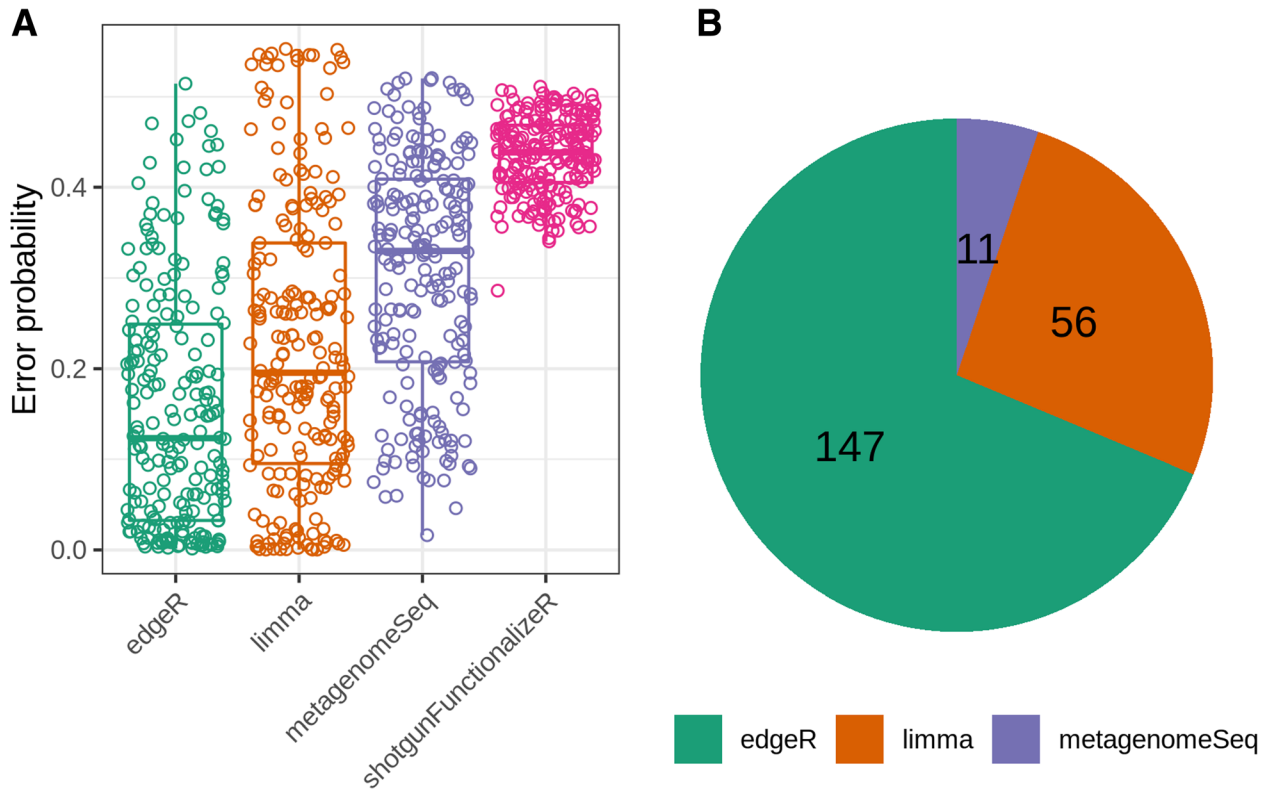

Fig. 5 A: Every dot represents the probability that a significant association is false. B: Among 214 simulations, edgeR had the lowest error probability in 147 simulations; limma had the lowest error probability in 56 simulations; metagenomeSeq had the lowest error probability in 11 simulations

this study ethnic-specific food groups were not included in this analysis and 67 food items were used for the analyses.

16S processing

We used the 16S ribosomal RNA (rRNA) sequencing data generated in a previous study based on the HELIUS cohort [3]. In short, the composition of fecal microbiota was profiled by sequencing the V4 region of the 16S rRNA gene on a MiSeq system. The 16S rRNA gene reads were processed on a mothur pipeline (version 1.39.5) [12]. The OTU clustering was done by using the vsearch (version 2.6) [13] and a phylogenetic tree was constructed by running FastTree 2.1 [14]. The details of the sequencing and bioinformatic pipelines were described in [3].

Statistical analyses

Our analysis is based on 1090 subjects who had both fecal microbiome and FFQ data. Following [1], here we removed OTUs with fewer than 10 reads in total, as well as OTUs which were present in fewer than 1% of samples. The final OTU table contains 1090 samples and 2073 OTUs. We used four widely used methods for sequencing data analysis to quantify the strength of the associations between dietary variables (x) and OTU counts (y). Because the large number of associations (67×2073), we used multidplyr R package (https://github.com/hadley/multidplyr) for parallel computation. The selected methods were as follows:

ShotgunFunctionalizeR is a popular R package used in microbiome research community, and based on the Poisson generalized linear model (implemented in glm function in R) [15]. We used the glm function with log(total counts) as offset to quantify associations between dietary variables and OTU counts.

Negative binomial model, also called gamma-Poisson model, is popular for statistical modeling of OTU count data [16, 17]. Phyloseq is a popular R package used by the microbiome research community [18]. The core of Phyloseq is based on another popular R package DESeq2, which is based on negative binomial model [19]. However, when the sample size is big (above 100), the computation becomes slow in DESeq2. Therefore, in this work we used another negative binomial based R package, edgeR [20]. The observed OTU count was modeled by a negative binomial distribution with two parameters, the mean and the dispersion. OTU specific

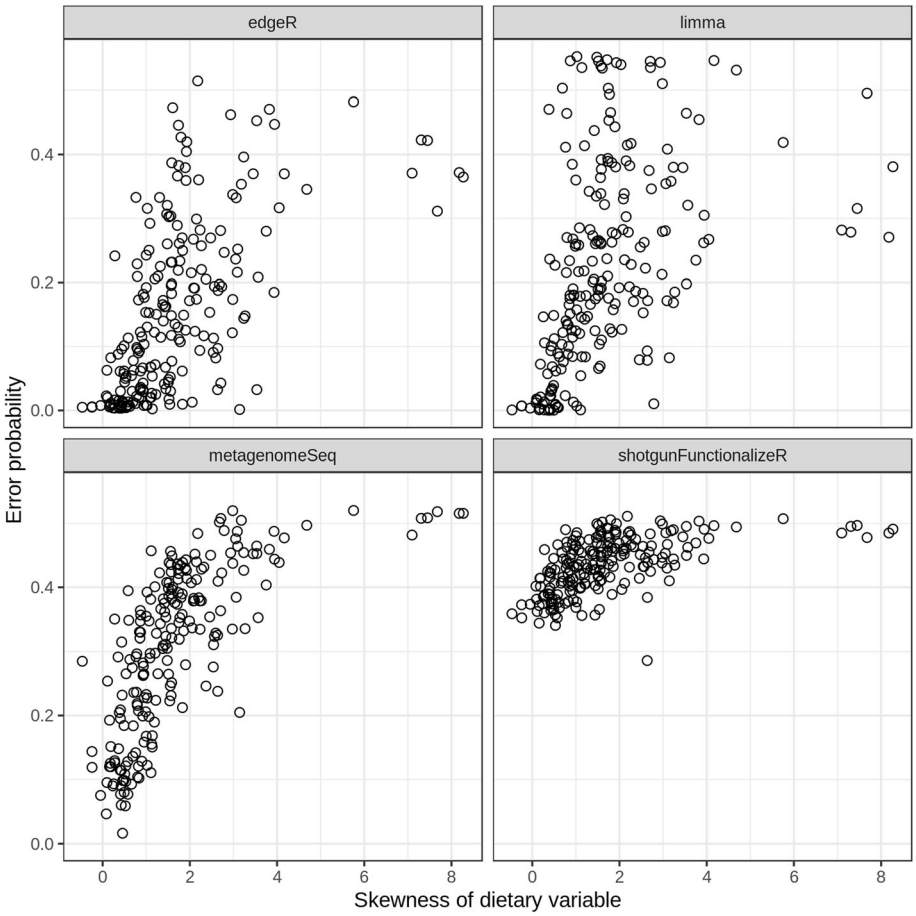

Fig. 6 Skewness of predictor variable influences false positive rate. Every circle represents a simulation

dispersion was estimated by running estimateDisp function implemented in the edgeR package [20, 21]. The associations between dietary variables and OTU counts were quantified by running glmFit function of the edgeR package [20]. The log(total counts) was used as offset.

In contrast to above methods modeling the counts with exact probabilistic distributions, others have advocated weighted linear regression analysis with precision weights derived from the mean variance relationship [22]. This approach has been implemented in the voom function of the popular R package limma [23]. The weighted linear regression was done to estimate the linear association between dietary variables and OTU counts with precision weights estimated by the voom and lmFit functions in the limma package [22, 23].

The last method, metagenomeSeq is also a popular R package used by microbiome research community [24]. It is based on the zero-inflated Gaussian model. This approach has been implemented in the fitZig function of the popular R package metagenomeSeq [24]. The cumulative sum scaling method was used to take care library size difference.

In a typical association study, the primary goal is to identify some candidate associations for future research. Therefore, regarding multiple testing we calculated false discovery rate (FDR). If an association had FDR value below 0.05, we considered it as a significant association. Since the research question is focused only on robustness of the statistical results and not on biological or epidemiological associations, we did not adjust for possible confounding or selection factors.

Simulation framework

We use y to represent the simulated microbiome data with n rows and J columns. Every column of y represents a microbe and every row of y represents a subject. Here, we simulated associations of a dietary variable, denoted as x, with gut microbiota. x is a vector of length n, and was randomly sampled from real FFQ data with replacement. The FFQ data was published in [2] and contained 214 dietary variables

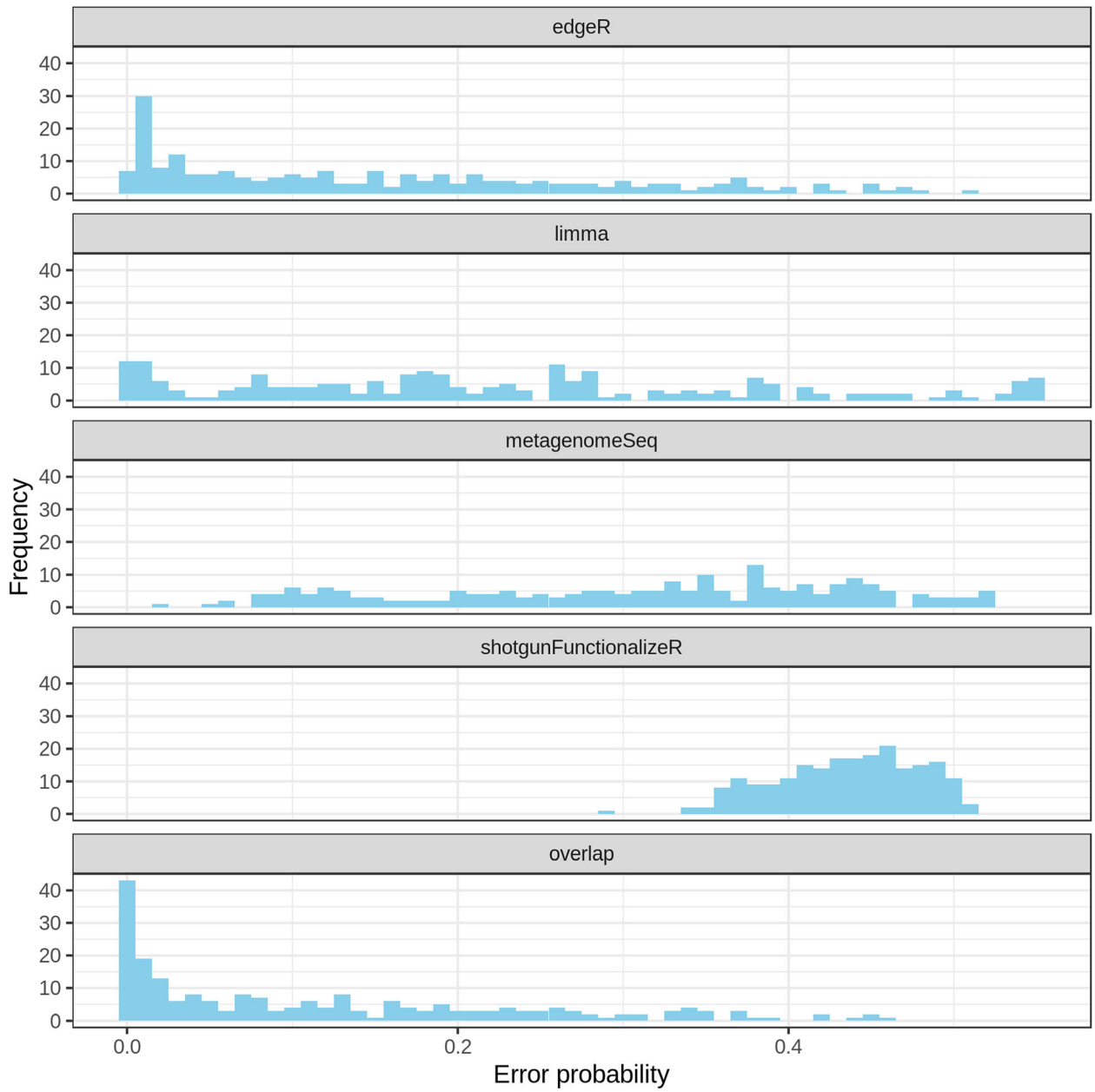

Fig. 7 Distribution of probabilities that a significant association is false in edgeR, limma, metagenomeSeq and shotgunFunctionalizeR. The "overlap" refers to the distribution of error probabilities of significant associations identified by all four methods

that were scaled to having mean 0 and standard deviation 1. For each simulated microbiome data set, we used one dietary variable and in total generated 214 simulated data sets. Our simulation framework included the steps below:

$$\eta[j] \sim Bernoulli(0.5) \quad (1)$$

$$\gamma[j] \sim T_7(0, 2.5) \quad (2)$$

$$\beta[j] = (1 - \eta[j]) \times 0 + \eta[j] \times \gamma[j] \quad (3)$$

$$\theta[i, 1:J] \sim Dirichlet(\pi[1:J]) \quad (4)$$

$$\alpha[i, 1:J] = logit(\theta[i, 1:J]) \quad (5)$$

$$logit(\mu[i, j]) = \alpha[i, j] + \beta[j] \times x[i] \quad (6)$$

$$N[i] \sim Lognormal(\mu_L, \sigma_L) \quad (7)$$

$$y[i, 1:J] \sim Multinomial(N[i], \mu[i, 1:J]) \quad (8)$$

Our HELIUS microbiome data set had 1090 subjects and the median sequencing depth was about 50,000. To mimic HELIUS data, we simulated the 16S microbiome

data sets, with each data set having 1000 subjects and mean of sequencing depth 50,000.

Spiked-in association

To introduce the spiked-in association between a dietary variable and microbe j, we need two variables $\eta[j]$ and $\gamma[j]$. The indicator variable, $\eta[j]$, indicates if a dietary variable influences the abundance of the microbe j. For microbe j, we randomly drew $\eta[j]$ from a Bernoulli distribution with parameter 0.5 (Eq. 1). $\gamma[j]$ represents the effect of the dietary variable on the abundance for OTU j, and was sampled from a t distribution with 7 degrees of freedom, location 0 and scale 2.5 [25] (Eq. 2). Then the true association between the diet and microbe j was captured by $\beta[j]$ defined in Eq. 3.

Dirichlet multinomial model

In the current study, we developed a Dirichlet multinomial model to generate 16S rRNA microbiome data.

In Eq. 4, the matrix θ has n rows and J columns. $\theta[i, j]$ corresponds to the baseline abundance level for the microbe j in subject i. For subject i, we randomly drew a vector of length J from a Dirichlet distribution. In Eq. 5, the parameter of the Dirichlet distribution π is a vector of length J. We used R package DirichletMultinomial [26] and the Human Microbiome Project 16S rRNA stool data [27] to estimate the π. In Eq. 6, the true microbe j proportion in subject i, $\mu[i, j]$ was modeled as a logistic regression of $x[i]$. Similar to [24], library size of subject i, $N[i]$, was randomly drawn from a lognormal distribution with mean μ_L and standard deviation σ_L (Eq. 7). μ_L is the logarithm of target sequencing depth (50000). We estimated $\sigma_L = 0.77$ based on the HMP stool 16S rRNA data set by using the fitdistr function implemented in the MASS package. Finally, for subject i, the observed microbe counts were randomly generated from a multinomial distribution (Eq. 8).

Performance evaluation

We evaluated the model performances based on metrics including true positive rate, false positive rate and error probability for identifying a significant association between microbe and dietary variable. They are calculated per simulation and defined as below:

$$\text{True positive rate} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{9}$$

$$\text{False positive rate} = \frac{\text{FP}}{\text{TN} + \text{FP}} \tag{10}$$

$$\text{Error probability} = \frac{FP}{\text{TP} + \text{FP}} \tag{11}$$

TP, FP, TN and FN refer to true positive, false positive, true negative and false negative, respectively. The indicator variable, $\eta[j]$, is in the definition of our spike-in associations. When $\eta[j] = 1$, the dietary variable x influences the abundance of the microbe j, otherwise $\eta[j] = 0$. Therefore, a true positive finding is defined as having a significant association between the dietary variable x and microbe j with FDR < 0.05 in case the true $\eta[j] = 1$. A false positive finding is defined as having a significant association between the dietary variable x and microbe j with FDR < 0.05 in case the true $\eta[j] = 0$. A true negative finding is defined as having a association between the dietary variable x and microbe j with FDR > 0.05 in case the true $\eta[j] = 0$. A false negative finding is defined as having a association between the dietary variable x and microbe j with FDR > 0.05 in case the true $\eta[j] = 1$. The error probability quantified the probability that a significant association is false. Here we did not use "false discovery rate" but used the term "error probability" in order to avoid confusion, because we also calculated the false discovery rate during analyses of associations between OTUs and dietary variables.

Abbreviations

FDR: False discovery rate; FFQs: Food frequency questionnaires; FN: False negative; FP: False positive; HELIUS: HEalthy Life in an Urban Setting; OTU: Operational taxonomic unit; TN: True negative; TP: True positive

Acknowledgements

Not applicable.

Funding

This work was supported by a personal ZONMW-VIDI grant 2013 [016.146.327] and a Dutch Heart Foundation CVON 2012 Grant (IN-CONTROL) [2012–03]. The funders had no role in the study design, the collection, analysis, and interpretation of data, the writing of the report, and the decision to submit the article for publication.

Availability of data and materials

The 16S rRNA gene sequences have been deposited at the European Genome-phenome Archive under study number EGAD00001004106. The FFQs data of this study are available from the study coordinator upon reasonable request. The 16S rRNA data and FFQs data, as well as the code used for simulation studies can be found at https://github.com/XiangZhangSC/HELIUS.

Consent to publish

Not applicable.

Authors' contributions

XZ performed the analysis and wrote the manuscript. MN generated the HELIUS microbiome data. MN, AKG and AHZ edited the manuscript. All authors have read and approved the manuscript.

Ethics approval and consent to participate

The HELIUS study was complied with all relevant ethical regulations and in accordance with the Declaration of Helsinki (6th, 7th revisions); it was approved by the Academic Medical Center (AMC) Medical Ethics Committee and all participants provided written inform ed. consent.

Competing interests

The authors declare that they have no competing interests.

Author details
[1]Department of Experimental Vascular Medicine, Amsterdam University Medical Center, Meibergdreef 9, Amsterdam, The Netherlands. [2]Department of Internal and Vascular Medicine, Amsterdam UMC, University of Amsterdam, Meibergdreef 9, Amsterdam, The Netherlands. [3]Department of Clinical Epidemiology, Biostatistics, and Bioinformatics, Amsterdam UMC, University of Amsterdam, Meibergdreef 9, Amsterdam, The Netherlands.

Received: 5 February 2019 Accepted: 24 April 2019
Published online: 09 May 2019

References

1. Duvallet C, Gibbons SM, Gurry T, Irizarry RA, Alm EJ. Meta-analysis of gut microbiome studies identifies disease-specific and shared responses. Nat Commun. 2017;8:1784. https://doi.org/10.1038/s41467-017-01973-8.
2. Wu GD, Chen J, Hoffmann C, Bittinger K, Chen Y-Y, Keilbaugh SA, et al. Linking long-term dietary patterns with gut microbial enterotypes. Science (New York, NY). 2011;334:105–8. https://doi.org/10.1126/science.1208344.
3. Deschasaux M, Bouter KE, Prodan A, Levin E, Groen AK, Herrema H, et al. Depicting the composition of gut microbiota in a population with varied ethnic origins but shared geography. Nat Med. 2018;24:1526–31. https://doi.org/10.1038/s41591-018-0160-1.
4. Xia Y, Sun J. Hypothesis testing and statistical analysis of microbiome. Genes & diseases. 2017;4:138–48. https://doi.org/10.1016/j.gendis.2017.06.001.
5. Thorsen J, Brejnrod A, Mortensen M, Rasmussen MA, Stokholm J, Al-Soud WA, et al. Large-scale benchmarking reveals false discoveries and count transformation sensitivity in 16S rRNA gene amplicon data analysis methods used in microbiome studies. Microbiome. 2016;4:62. https://doi.org/10.1186/s40168-016-0208-8.
6. Jonsson V, Österlund T, Nerman O, Kristiansson E. Statistical evaluation of methods for identification of differentially abundant genes in comparative metagenomics. BMC Genomics. 2016;17:78. https://doi.org/10.1186/s12864-016-2386-y.
7. Stronks K, Snijder MB, Peters RJG, Prins M, Schene AH, Zwinderman AH. Unravelling the impact of ethnicity on health in Europe: the HELIUS study. BMC Public Health. 2013;13:402.
8. Vermeulen E, Stronks K, Visser M, Brouwer IA, Snijder MB, Mocking RJT, et al. Dietary pattern derived by reduced rank regression and depressive symptoms in a multi-ethnic population: the HELIUS study. Eur J Clin Nutr. 2017;71:987–94. https://doi.org/10.1038/ejcn.2017.61.
9. Pereira MB, Wallroth M, Jonsson V, Kristiansson E. Comparison of normalization methods for the analysis of metagenomic gene abundance data. BMC Genomics. 2018;19:274. https://doi.org/10.1186/s12864-018-4637-6.
10. Dekker LH, Snijder MB, Beukers MH, de VJHM, Brants HAM, de BEJ, et al. A prospective cohort study of dietary patterns of non-western migrants in the Netherlands in relation to risk factors for cardiovascular diseases: HELIUS-dietary patterns. BMC Public Health. 2011;11:441.
11. Beukers MH, Dekker LH, de BEJ, Perenboom CWM, Meijboom S, Nicolaou M, et al. Development of the HELIUS food frequency questionnaires: ethnic-specific questionnaires to assess the diet of a multiethnic population in the Netherlands. Eur J Clin Nutr. 2015;69:579–84.
12. Schloss PD, Westcott SL, Ryabin T, Hall JR, Hartmann M, Hollister EB, et al. Introducing mothur: open-source, platform-independent, community-supported software for describing and comparing microbial communities. Appl Environ Microbiol. 2009;75:7537–41. https://doi.org/10.1128/AEM.01541-09.
13. Rognes T, Flouri T, Nichols B, Quince C, Mahé F. VSEARCH: a versatile open source tool for metagenomics. PeerJ. 2016;4:e2584. https://doi.org/10.7717/peerj.2584.
14. Price MN, Dehal PS, Arkin AP. FastTree 2–approximately maximum-likelihood trees for large alignments. PLoS One. 2010;5:e9490. https://doi.org/10.1371/journal.pone.0009490.
15. Kristiansson E, Hugenholtz P, Dalevi D. ShotgunFunctionalizeR: an R-package for functional comparison of metagenomes. Bioinformatics (Oxford, England). 2009;25:2737–8. https://doi.org/10.1093/bioinformatics/btp508.
16. McMurdie PJ, Holmes S. Waste not, want not: why rarefying microbiome data is inadmissible. PLoS Comput Biol. 2014;10:e1003531. https://doi.org/10.1371/journal.pcbi.1003531.
17. Zhang X, Mallick H, Tang Z, Zhang L, Cui X, Benson AK, et al. Negative binomial mixed models for analyzing microbiome count data. BMC Bioinformatics. 2017;18:4.
18. McMurdie PJ, Holmes S. Phyloseq: an R package for reproducible interactive analysis and graphics of microbiome census data. PLoS One. 2013;8:e61217.
19. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol. 2014;15:550. https://doi.org/10.1186/s13059-014-0550-8.
20. Robinson MD, McCarthy DJ, Smyth GK. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics. 2010;26:139–40.
21. Chen Y, Lun ATL, Smyth GK. Differential expression analysis of complex RNA-seq experiments using edgeR. In: Datta S, Nettleton D, editors. Statistical analysis of next generation sequencing data. Cham: Springer International Publishing; 2014. p. 51–74. https://doi.org/10.1007/978-3-319-07212-8_3.
22. Law CW, Chen Y, Shi W, Smyth GK. Voom: precision weights unlock linear model analysis tools for RNA-seq read counts. Genome Biol. 2014;15:R29. https://doi.org/10.1186/gb-2014-15-2-r29.
23. Ritchie ME, Phipson B, Wu D, Hu Y, Law CW, Shi W, et al. Limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic Acids Res. 2015;43:e47. https://doi.org/10.1093/nar/gkv007.
24. Paulson JN, Stine OC, Bravo HC, Pop M. Differential abundance analysis for microbial marker-gene surveys. Nat Methods. 2013;10:1200–2. https://doi.org/10.1038/nmeth.2658.
25. Gelman A, Jakulin A, Pittau MG, Su Y-S. A weakly informative default prior distribution for logistic and other regression models. The Annals of Applied Statistics. 2008;2:1360–83. https://doi.org/10.1214/08-AOAS191.
26. Holmes I, Harris K, Quince C. Dirichlet multinomial mixtures: generative models for microbial metagenomics. PLoS One. 2012;7:e30126.
27. Schiffer L, Azhar R, Shepherd L, Ramos M, Geistlinger L, Huttenhower C, et al. HMP16SData: Efficient Access to the Human Microbiome Project through Bioconductor. bioRxiv. 2018. http://biorxiv.org/content/early/2018/08/29/299115.abstract.